INFORMATION AND CONTROL 15, 527–528 (1969)

A Note on the Determination of a Discrete Probability Distribution from Known Marginals

BUSH JONES

Mail Point 411, Martin Marietta Corp., P. O. Box 5837, Orlando, Florida 32805

AND

ORAN BRIGHAM

National Security Agency

A theorem that gives the minimum number of values of a discrete probability distribution, which must be known to determine the distribution from a known set of marginals, is presented.

This note is concerned with conditions for the exact determination of a discrete probability distribution $P(X_1, X_2 \cdots, X_n)$, denoted in this note by P, when some of its marginal distributions are known. Brown (1959) treated the problem of approximating a probability distribution from its marginals. The usual properties of a joint probability distribution and its marginals are assumed. A marginal distribution is denoted as p_i, a particular but arbitrary value of a marginal p_i is denoted by q_m, and a particular but arbitrary value of P by P^j. The equation

$$q_m = \sum_j A_{mj} P^j, \tag{1}$$

where each A_{mj} is either 0 or 1, follows from the properties of a marginal distribution and is referred to as a constraint equation. For a set of marginals $\{p_i\}$, $\{q_m\}$ denotes the set of all values of the marginals. If there are l values in $\{q_m\}$, then the matrix constraint equation

$$q = AP \tag{2}$$

represents l linear equations in 2^n unknowns. Before stating the theorem which gives the number of independent equations contained in (2), the following definition must be made.

DEFINITION. Given a set of marginals $\{p_i\}$, the set of marginals D "defined by" $\{p_i\}$ consists of all marginals defined on any sequence or subsequence of the sequences of binary variables that the p_i are defined on.

EXAMPLE. Given $\{p_i\} = \{p(X_1, X_2), p(X_2, X_3)\}$,
then

$$D = \{p(X_1), p(X_2), p(X_3), p(X_1, X_2), p(X_2), X_3)\}.$$

THEOREM. *If $\{p_i\}$ is a set of known marginal distributions of an nth order discrete probability distribution P, and K is the number of members in the set D defined by $\{p_i\}$, then $2^n - (K + 1)$ is the minimum number of values of P that must be known to completely determine P.*

Proof. Consider the set consisting of the value 1 together with the set of values obtained by taking one arbitrary value of each marginal in the set D. The following statements are easily shown.

(a) The constraint equations for this set are independent.

(b) Any subset that is a subset of the values of marginals in D and has the property that any value of any marginal in D can be determined from it, has the same number of independent constraint equations as the above set.

Now $\{q_m\}$, the set of all values of the marginals in $\{p_i\}$, satisfies the property described in (b), and the Theorem is thus implied by (b).

It is worth noting that a set of independent equations can be formed to solve for P. These are the constraint equations for the set defined at the onset of the above proof.

RECEIVED: May 28, 1969; REVISED: September 2, 1969.

REFERENCE

1. DAVID BROWN, A note on approximations to discrete probability distributions, *Information and Control* **2** (1959), 386.